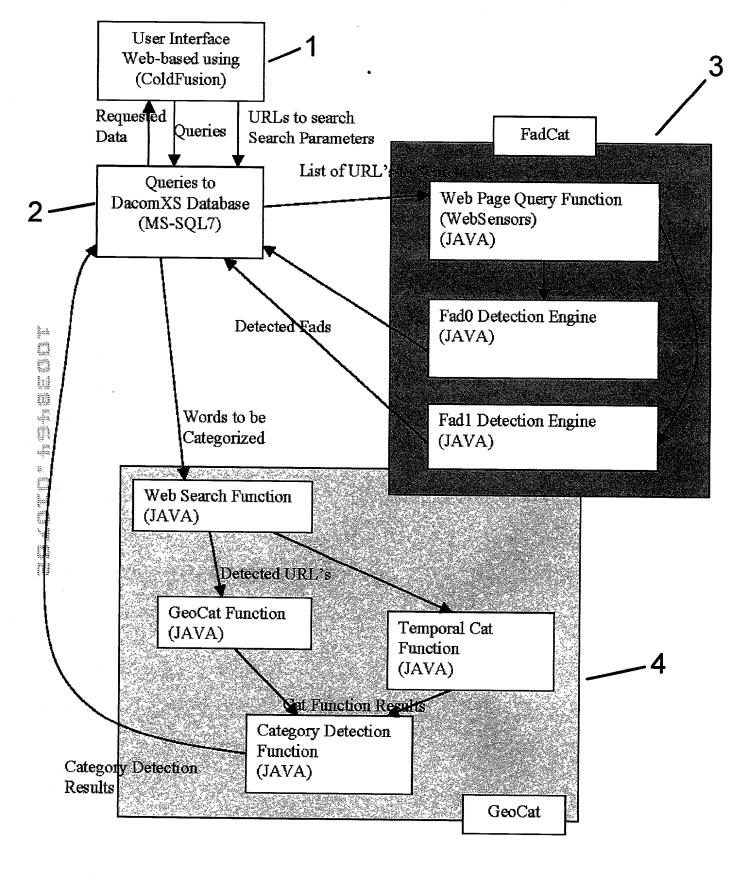
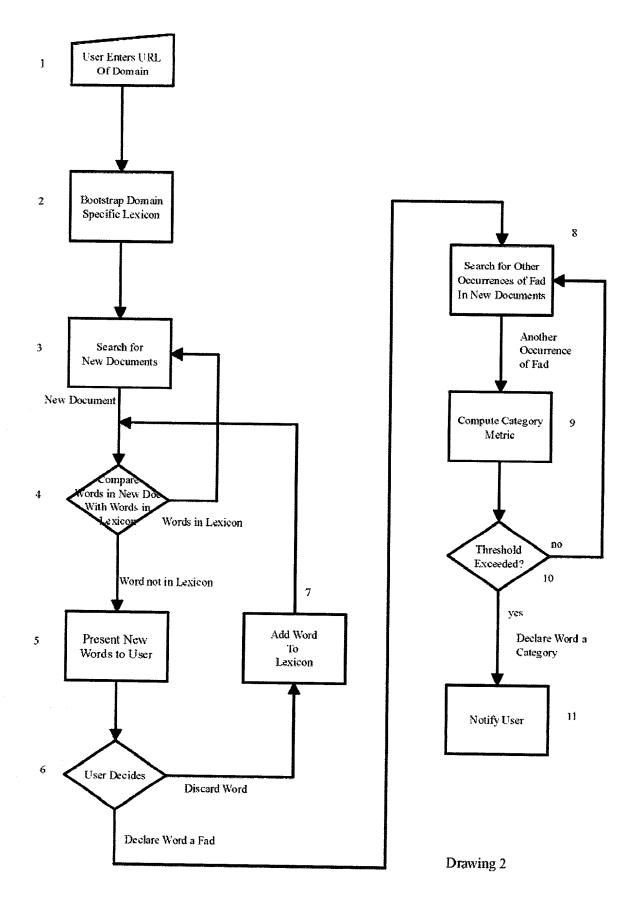
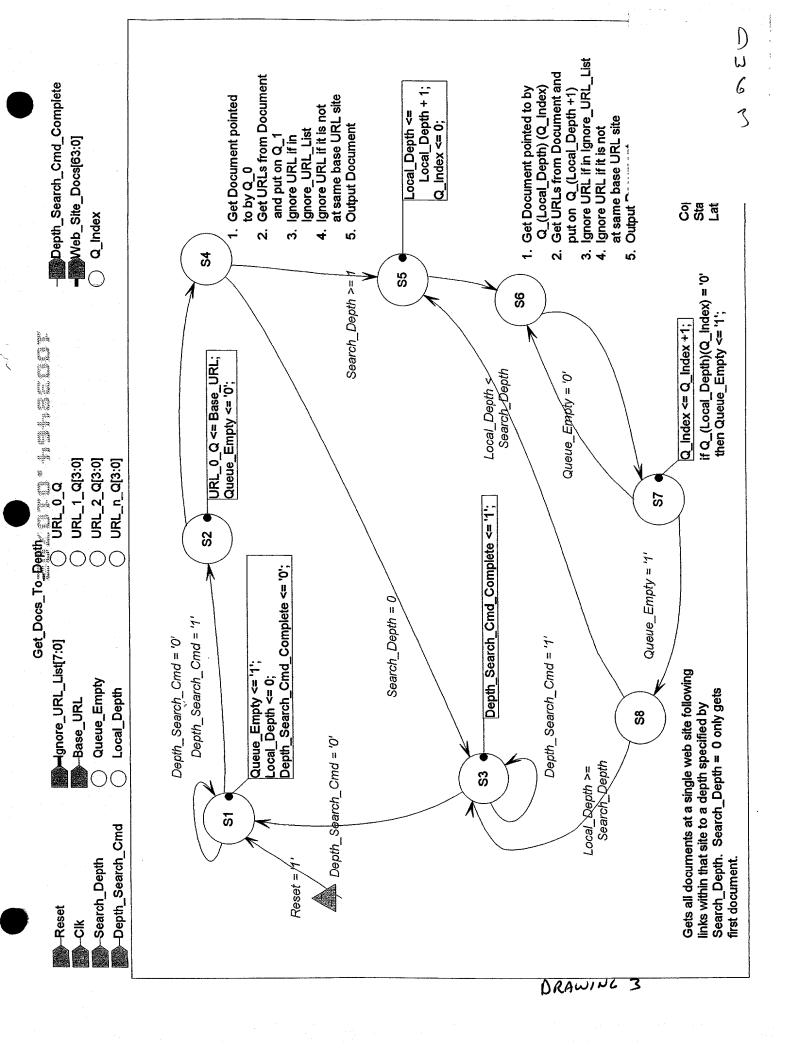
DRAWINGS







0000 Copyright: K. J. Hintz, 2000 Start Date: 4/3/00 Latest Update: 4/4/00 Gets all documents from a web site and from all web sites pointed to by that web site except for URLs in the Ignore_URL_List (sites like yahoo, lycos, altavista, etc.) 2. Get URLs from Document and put on Q
3. Ignore URL if on Ignore_URL_List
3. Output Document Get Document pointed to by top of Q f URL_Queue (URL_Index) = '0' then Queue_Empty <= '1'; JRL_Index <= URL_Index + 1 URL_Queue (URL_Index) <= Base_URL; Queue_Empty <= '0'; end if; \$ **S**2 82 $Queue_Empt_{N} = 0$ Web Site Cmd Complete <= '1'; URL_Queue <= "00000000000000000000"; Web_Site_Cmd_Complete <= '0'; Get Web Site Cmd = '1' Queue_Empty = '1' Get_Web_Site_Cmd = '0' Queue_Empty <= '1'; JRL_Index <= 0; Get_Web_Site_Cmd = '0' Get_Web_Site_Cmd = '1'

Set_Web_Site_Cmd = '1' 83 S Reset = 11 Sreg0

Web_Site_Cmd_Complete

Web_Site_Docs[63:0]

O URL Queue 15:0]

Get_All_Docs

Hgnore_URL_List[7:0]

Base_URL

-Get_Web_Site_Cmd

Cik

Queue_Empty

URL_Index

Copyright: K. J. Hintz, 2000 Start Date: 4/3/00 to by URL_Q(Q_Index)
Get URLs from Document If hash value has changed, Compare hash value of **Get Document pointed** at same base URL site **Output Document and** document to last hash Latest Update: 4/4/00 Ignore_URL_List Ignore_URL_if it is not and put on Queue Ignore URL if in update hash list if URL_Q(Q_Index) = '0' then Queue_Empty <= '1'; Q_Index <= Q_Index +1; က Š. ဖ \$ Queue_Empty = /0 URL_Q(Q_Index) <= Base_URL; Queue_Empty <= '0'; 27 Queue_Empty = '1' New_Docs_Cmd_Complete <= '1'; **8**5 Queue_Empty <= '1';
New_Docs_Cmd_Complete <= '0';
Q_Index <= 0; New_Docs_Cmd = '1' New_Docs_Cmd = '1' Gets only those documents from one site which Uses a list of document "hash" values stored by URL for later comparison have changed since the last time they were New_Docs_Cmd = '0' $New_Dods_Cmd = '0'$ င္ယ က Reset =/ accessed.

DRAWING

0 ا ا

New_Docs_Cmd_Complete

Get_New_Docs

-dgnore_URL_List[7:0]

-Reset

Hash_List[7:0]

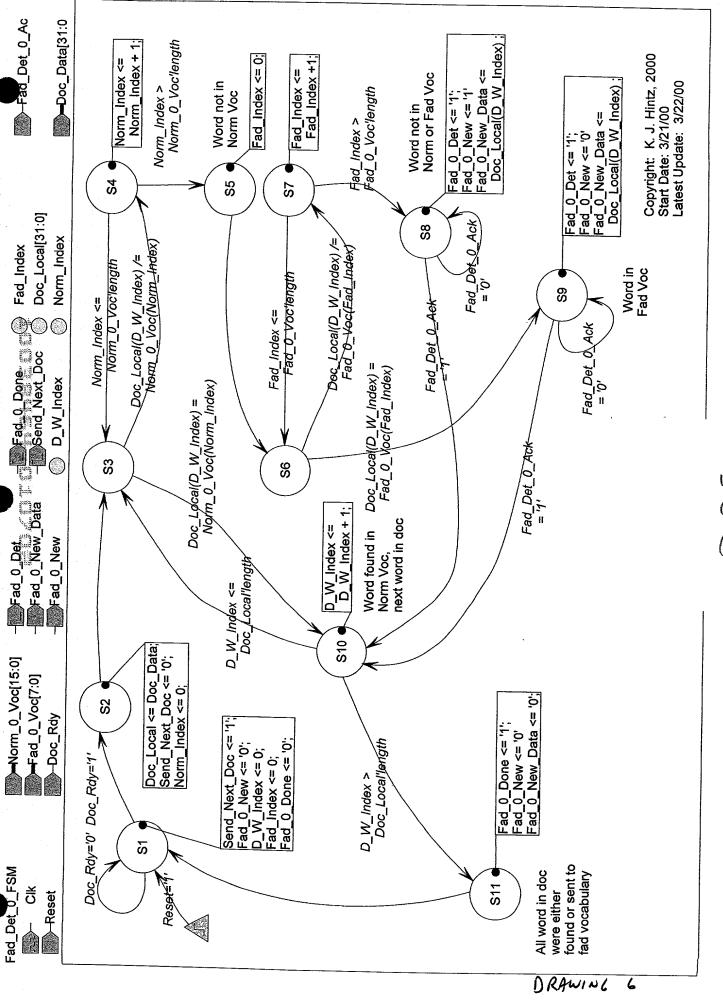
New Docs Cmd

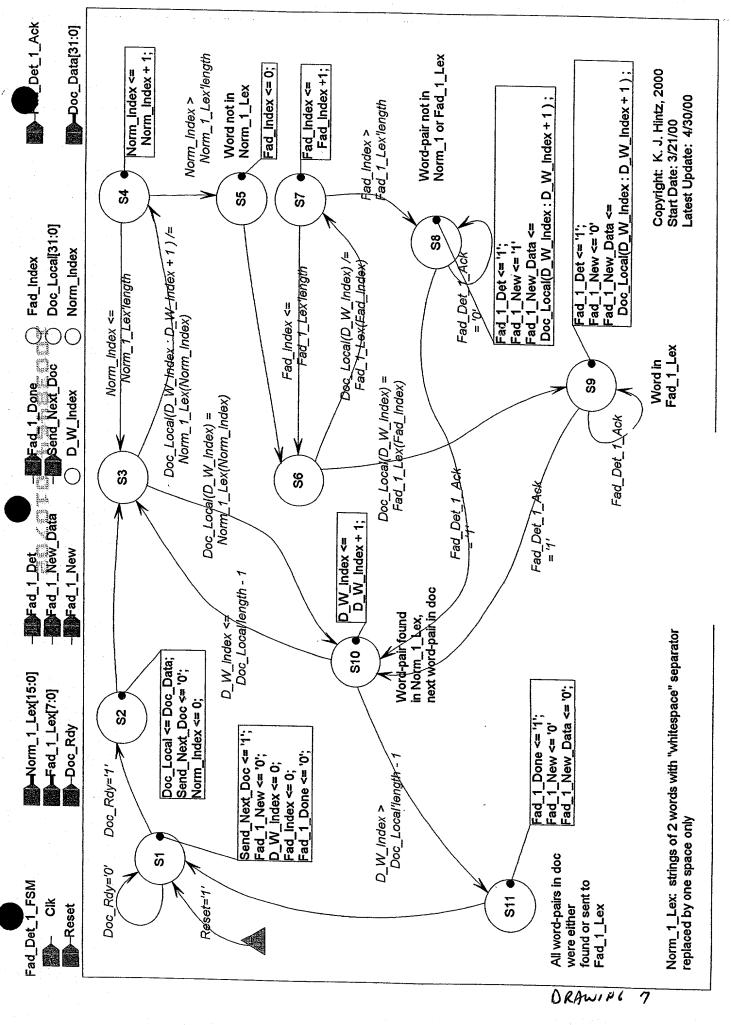
-Base_URL

Queue_Empty

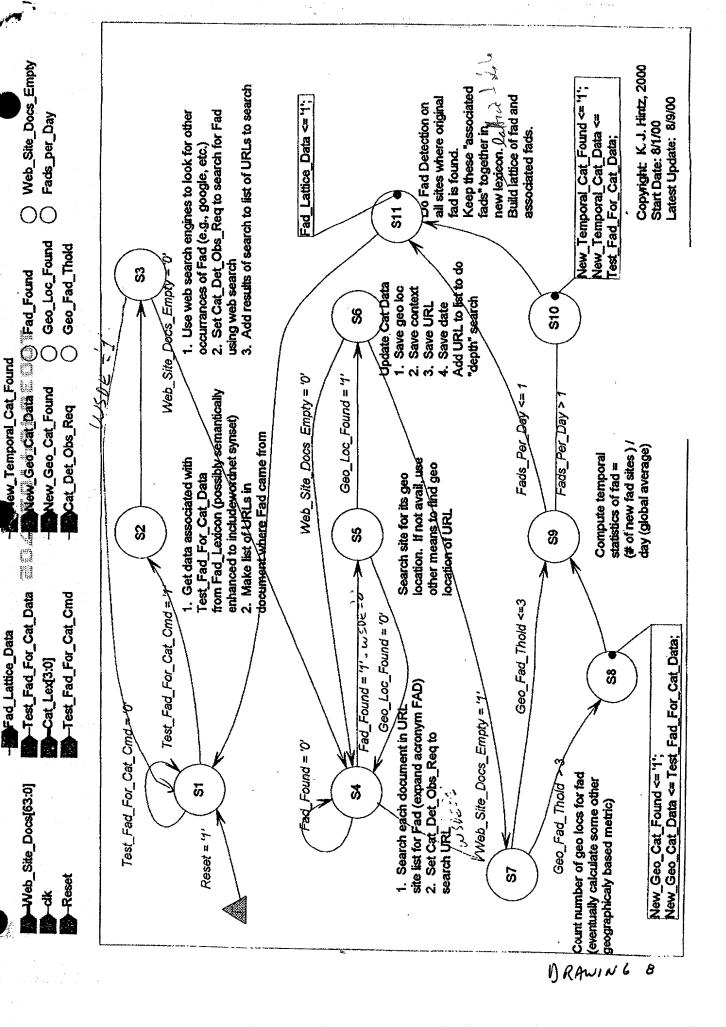
Q_Index

- Neb_Site_Docs[63:0]





F-67



✓ Temporal Cat Data

CAT_DET_FSM